The Use of Data Mining to Investigate a Possible Quality Problem with Ultrasensitive HIV Viral Load Data at a Large Reference Laboratory

Eileen Koski, M.Phil.[1], Peter N.R. Heseltine, M.D., FACP[2], Ron M. Kagan, Ph.D.[2], Ann E. Maddo, B.S., MT(ASCP)[1] and Jake Geller, Ph.D.[1]

[1]Quest Diagnostics Incorporated, Advanced Diagnostics IT, Teterboro, NJ

[2]Quest Diagnostics Incorporated, Nichols Institute, San Juan Capistrano, CA

Abstract: Suppression of HIV viral load to <50 copies/mL, the lower limit of detection for the ultrasensitive assays, has been shown to correlate with favorable clinical outcome. Patients periodically exhibited transient or sustained low-level viremia based on this test. In order to investigate a possible quality concern, we used our corporate data warehouse to examine the patterns in our data over time as well as across geographic regions.

Problem: There was a concern that transient or low-level viremia observed in patients who had previously maintained an undetectable viral load might represent a quality problem with our ultrasensitive assays. Historically, gathering the data to investigate such a problem required extracting data from archives at each site and then attempting to compile a data set from source data in varying formats from different systems. Quest Diagnostics has, however, established a corporate data warehouse of order and result data from all of our local, regional and esoteric laboratories in one repository.

Project: Using our corporate data warehouse, we were able to rapidly identify and retrieve results for clinical samples submitted for HIV RNA quantitative assays (ultrasensitive RT-PCR and bDNA 3.0) from all our laboratories. Data was retrieved for 168,383 usPCR results and for 103,556 bDNA results collected over a 16-month period from Oct. 2001 to Jan. 2003.

The raw data retrieved from the data warehouse was segmented for use in different ways for different analyses. For example, data submitted to us directly from a physician typically includes geo-coding information in the form of a physician and patient zip code, while data submitted by a hospital or other referring laboratory does not. In addition, only six of our laboratories actually perform this test, even though all of our laboratories transmit their results to the data warehouse whether performed by them or elsewhere within our company. Additional steps were therefore required to eliminate duplicate records since each lab that touches a specimen assigns their own specimen identifiers and date of service making it impossible for us to use any simple strategy such as retaining only unique records.

The final data set was analyzed, including analyses of variance across business units.

Conclusions: We were able to determine that low-level viremia is consistently detected in a large, geographically dispersed, clinical data set by two quantitative HIV-1 RNA technologies, indicating that low-level viremia is not likely to be an anomaly due to technical variability. From a data mining perspective, the project established the value of our corporate data warehouse to support rapid examination and analysis of issues related to our clinical data.